学校编码：10384　　分类号_____密级_____
学号：200410013　　UDC_____

硕 士 学 位 论 文

多分类有序变量间距差异的统计分析与实际应用

The Statistical Analysis and Its Application for the Interval Differences between Categories of Ordinal Data

陈民恳

指导教师姓名：朱 建 平 教授
专 业 名 称：统 计 学
论文提交日期：2007 年 3 月
论文答辩时间：2007 年 月
学位授予日期：2007 年 月

答辩委员会主席：
评 阅 人：

2007 年 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（ ），在　　年解密后适用本授权书。

2、不保密（ ）

（请在以上相应括号内打“ √ ”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

内容摘要

定性数据的统计分析是当前的热点，也是难点问题。定性数据常以多分类有序变量的形式出现，尤其是在市场调查和社会科学研究所收集的数据中。多分类有序变量是指分类数大于等于3，且类别之间存在序次关系的响应变量。在对此类资料进行统计分析的过程中，我们发现，有序变量的“类间距”并不相等，也就是各类型之间的稀疏程度并不是均匀的。例如，人们对一个事物的评价从“很不喜欢”到“不喜欢”，再到“喜欢”、“很喜欢”，它们层级之间的差距通常是不同的，而一般的数据分析方法却将其作等距对待，这样的处理往往是粗糙而不精确的。

有关有序变量间距差异的研究在国外的文献中略有提及但并不深入，且往往集中在对有序变量的赋值研究上，而国内更是鲜有人涉及。正是针对此等研究现状，本文尝试着从统计学的角度对该问题进行详细、系统地论证和分析，并在实际工作中加以运用。全文主要分成四个部分：

第一部分：第一章。回顾了近半个多世纪来，国内外专家学者在有序变量领域所作的相关研究和成果。有序变量的统计分析理论在国外已发展地较为成熟，而国内则起步较晚，发展也较为落后，还没有形成国人自身的体系。

第二部分：第二、三章。针对 Likert 类型尺度，提出了衡量间距差异的统计量和统计检验方法，在此基础上，改进了累积 logistic 回归模型，并结合实例加以论证（创新点一）。对于有序变量，若经过检验发现确实存在间距差异，则应引入虚拟工具变量对累积 logistic 模型加以修正，以提高模型的精确度。

第三部分：第四章。针对程度等级变量，提出了用秩分析的方法对有序变量的间距差异进行了界定和量化（创新点二），并在此基础上，引入多元统计分析方法中的聚类分析，对量化后的资料进行了实证应用，取得了和前人相近的结论，提高了效率。

第四部分：第五章。针对分组有序变量，以我国地区经济差距为例，在实证分析中介绍了该类型有序变量间距差异的界定和统计检验方法（创新点三），并运用计量建模分析对我国地区经济差距提出了自己的见解。

关键词：有序变量；间距差异；统计分析

Abstract

Statistical analysis of qualitative data is one of the hottest but difficult issues to study recently. Qualitative data always comes up as ordered categorical variables, especially in the data set collected by marketing and social science researching. Ordered categorical variables means the response variables that has more than three categories between which are ordered. We found that the interval between each two categories was different when we on the process of such kind of data. For example, when people choose to evaluate a thing, their attitude from “strongly unlike” to “unlike”, and then from “like” to “strongly like” should be asymmetrically distributed. Most of us always ignore it and treat it has equal intervals, which may lead to draw an inexact conclusion.

Though this subject have been mentioned among some research papers overseas, it hasn’t gone thoroughly and mostly focused on how to evaluate ordinal data, let alone within our nation. Under such condition, this paper attempt to analyze the intervals between categories of ordinal data elaborately, systematically and practically from the statistic point of view. The paper is divided into four parts as follows:

Part one: chapter one. Take a review on the foreign and domestic documents concerning ordinal data through half a century. It turns out that the theory of ordinal data analysis has been fully developed overseas compared to our country where hasn’t established its own theory in this field.

Part two: chapter two and three. Put forward the statistic and test method to evaluate the differences between categories of ordinal data as Likert type. On this basis, we make an emendation on cumulative logistic model and demonstrate it in application (innovation point one). For Likert measurement, if there exists interval differences, one should introduce instrument dummy variable to modify cumulative logistic model so as to advance its precision.

Part three: chapter four. Propose the method of rank analysis to evaluate the differences between categories of ordinal data (innovation point two). On this basis, apply the cluster analysis to the quantification of the ordinal data and obtain a similar

conclusion of predecessor's, but enhance the efficiency.

Part four: chapter five. Take the regional discrepancies of our nation as the case; present the method of evaluating and testing the interval differences between grouped ordinal data (innovation point three). Then apply econometrical models to analyze the regional discrepancies and put forward my own understanding on this issue.

Keywords：ordered categorical variables；interval differences；statistical analysis

目 录

Contents

前 言

一、选题意义

定性数据的统计分析是当前的热点，也是难点问题。定性数据常以多分类有序变量的形式出现，尤其是在市场调查、医学和社会科学研究所收集的数据中。多分类有序变量是指分类数大于等于3，且类别之间存在序次关系的响应变量。有序变量共分为三种类型：纯有序变量（pure），如等级变量；弱测量尺度（imperfect scale measurement），如 Likert 类型尺度；分组变量（grouped variables），如按收入高低分组得到的变量（Brendan Halpin，2002）。一般而言，有序变量类别之间的距离并不固定，也就是各类型之间的稀疏程度并不均匀（Winship，Mare，1984；J. Scott Long，2002；朱建平，2005）。例如，人们对一个事物的评价从“很不喜欢”到“不喜欢”，再到“喜欢”、“很喜欢”，它们层级之间的差距往往是不同的，而一般的数据分析方法却将其作等距对待，这样的处理往往是粗糙而不精确的。

有关有序变量间距差异的研究在国外的文献中略有提及但并不深入，且往往集中在对有序变量的赋值研究上（Gautam，Kimeldorf 等，1996；Singer，Poleto 等，2004），而国内更是鲜有人涉及（张晋昕、李河，2005；丁元林，孔丹莉，2005）。正是针对此种研究现状，本文尝试着从统计学的角度对该问题进行详细、系统地论证和分析，并将有关方法在实际工作中加以运用。

二、研究的内容和方法

本着系统、翔实的原则，笔者分别对三种类型的有序变量的间距差异进行了分析和论证。首先就 Likert 类型尺度，提出了衡量间距差异的统计量，并对间距差异作出统计意义上的检验分析。在此基础上，引入工具虚拟变量对累积 logistic 回归模型进行了改进，并结合调查数据加以运用；随后，笔者针对程度等级变量，运用秩分析的方法对间距差异作出了界定，对界定后的资料进行量化处理，并进一步引入聚类分析，通过实例应用取得了良好的效果。最后，笔者以我国地区经济差距分析为例，在实证分析中探讨了分组有序变量的间距差异的统计分析方法。在此框架内，综合运用多元统计方法、统计检验和计量建模分析对我们地区经济差距的现状进行了研究。

第一章　多分类有序变量的统计分析方法概述

第一节　国外文献综述

一、多分类有序变量的建模分析

对多分类有序变量的分析，要考虑到类别之间的有序性。最初，人们简单地取各类别的组中点为代表值，将有序变量视为连续性资料，然后取对数做估计。如此，在回归分析中，将有序变量转化为连续变量的做法将产生误导之结果（Winship and Mare，1984）。因此，对有序变量的建模分析，应采用符合该变量特性的统计分析方法。

（一）有序变量的 Logistic 回归模型

假定有序变量 y 取 c 个不同的值 $1,2,\mathrm{L}\ ,c$，并且序和数字的大小相同，即 y 取的值就是它的秩（次序），记

$$\pi_i = P(y=i) \quad i=1,2,\mathrm{L}\ ,c$$

这时可以引入不同概率的比，取对数后相应的意义也不同（张尧庭，1991）。

1、累积（Cumulative）logit 变换

$$L_j = \ln\frac{\sum_{i=j+1}^{c}\pi_i}{\sum_{i=1}^{j}\pi_i} = \ln\frac{P(y\geq j+1)}{P(y\leq j)} = \ln\frac{P(y\geq j+1)}{1-P(y\geq j+1)},\quad j=1,2,\mathrm{L}\ ,c-1$$

这是依次将 c 类合并为两类，把两类作 logistic 回归，进行分析和比较。

2、相继的（Continuation）logit 变换

$$L_j = \ln\frac{\pi_{j+1}}{\sum_{i=1}^{j}\pi_i},\quad j=1,2,\mathrm{L}\ ,c-1$$

这实际上是条件 logistic 的回归，因为

$$\ln\frac{\pi_{j+1}}{\sum_{i=1}^{j}\pi_i} = \ln\frac{P(y=j+1)/P(y\leq j+1)}{P(y\leq j)/P(y\leq j+1)} = \ln\frac{P(y=j+1\,|\,y\leq j+1)}{1-P(y=j+1\,|\,y\leq j+1)}$$

它反映了单独考虑前面的 $j+1$类时，第 $j+1$类的条件概率受自变量因素的影响如何。

3、相邻的（Adjacent）logit 变换

$$L_j = \ln \frac{\pi_{j+1}}{\pi_j}, \quad j = 1,2,\mathrm{L}\ ,c-1$$

它也是一种条件概率的 logit 变换。

累积比率 logistic 模型是最为常用的有序变量回归模型，它假设存在一个不可观测的潜在连续变量 y_i^*，通过对 y_i^* 划分为不同的类别得到有序变量 y。该模型同时应用 $c-1$个累积概率，并且假设预报因子（predictors）对每个累积概率具有相同的影响，因而，又被称为比例优势模型（proportional odds model）(McCullagh，1980；Agresti，1999)。因此，在应用该模型时，事先需要对成比例发生比假设条件（proportional odds assumption）进行检验，如果该条件不成立，则需要选用其他模型进行分析。Bender and Grouven（1998）就曾针对此情形，探讨了用二元 logistic 回归模型代替累积 logistic 回归模型进行分析的情形。

（二）有序 Probit 模型

Probit 模型是 logit 模型的姐妹篇，取决于误差项的分布。若假设误差项服从对数分布则为 logit 模型，服从正态分布时即为 probit 模型。

对于多分类有序变量 Y_i，有如下模型形式：

$$Y_i = j \quad u_{i,j-1} < Y_i^* \le u_{i,j} \quad j = 1,2,\mathrm{L}\ ,m$$

潜变量（latent variable）Y_i^* 设为社会经济变量向量 X 的函数：

$$Y_i^* = \beta' \mathrm{X} + \varepsilon_i \quad \varepsilon_i :\ N(0,1)$$

由于假设误差项服从正态分布，则观测到有序变量 Y_i 取某一特定值的概率为：

$$P_{ij} = P(Y_i = j) = \phi(u_{ij} - \beta' \mathrm{X}) - \phi(u_{ij-1} - \beta' \mathrm{X})$$

其中，$\phi(g)$ 为累积标准正态分布函数，模型可采用最大似然法进行估计。

（三）联合模型（Association Models）

Logit 模型同一般回归模型一样，将变量区分为反应变量和解释变量，而联

合模型则将所有的变量不加区分地等同对待，重在描述变量之间的关联性。

对于$r\times c$列联表资料，每个单元的频数$\{n_{ij}\}$的期望值为$\{\mu_{ij}\}$，则可建立联合模型（Goodman，1985）：

$$\log\mu_{ij}=\lambda+\lambda_i^x+\lambda_j^y+\sum_{k=1}^{M}\beta_k u_{ik}v_{jk}$$

这里，$M\le\min(r-1,c-1)$，通常情况下，取$M=1$。双线性联合模型（linear-by-linear）将行得分（row scores）$\{\mu_{i1}\}$和列得分（column scores）$\{\mu_{j1}\}$都固定为常数；而行效应模型则将行得分视为可变参数而固定列得分；列效应模型而固定行得分而视列得分为可变参数。双效应模型（RC model）将行得分和列得分都视为可变参数，在这种情况下，联合模型不再是对数线性的，它的似然估计将更为复杂（Haberman，1995）。

（四）纯有序变量回归模型

针对现行有序变量模型或者忽略变量间的有序特性，或者有赖于潜在连续变量的假定的缺陷，Torra，Domingo-Ferrer（2006）等提出了基于最小二乘法的纯有序回归模型（regression model without underlying continuous variables）。

假定建立一个二维表，一维代表目标集，记为$O=(o_1,o_2,\text{L}\ ,o_M)$，另一维代表变量集，记为$V=(V_0,V_1,V_2,\text{L}\ ,V_N)$，则该二维表可建立函数形式如下：

$$V:O\to D(V_0)\times D(V_1)\times D(V_2)\times\text{L}\ \times D(V_N)$$

其中，$D(V_i)$表示变量V_i的取值范围。

为简单起见，选择某一有序变量为目标变量$V_0(o)$，建立该目标变量对其他有序变量的线性回归模型：

$$\hat{V}_0(o)=\hat{\beta}_1V_1(o)+\hat{\beta}_2V_2(o)+\text{L}\ +\hat{\beta}_NV_N(o)$$

运用最小二乘法可求解得到参数估计向量：

$$\hat{\beta}=(X^TX)^{-1}X^TV_0$$

其中，$\hat{\beta}=\{\hat{\beta}_1,\hat{\beta}_2,\text{L}\ ,\hat{\beta}_N\}$，$V_0=\{V_0(o_1),V_0(o_2),\text{L}\ ,V_0(o_M)\}$，$X=\{V_1,V_2,\text{L}\ ,V_N\}$。